Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.048270

ARTICLE

Leveraging User-Generated Comments and Fused BiLSTM Models to Detect and Predict Issues with Mobile Apps

Wael M. S. Yafooz* **and Abdullah Alsaeedi**

Department of Computer Science, College of Computer Science and Engineering, Taibah University, Medina, 42353, Saudi Arabia

*Corresponding Author: Wael M. S. Yafooz. Email: wyafooz@taibahu.edu.sa

Received: 02 December 2023 Accepted: 26 February 2024 Published: 25 April 2024

ABSTRACT

In the last decade, technical advancements and faster Internet speeds have also led to an increasing number of mobile devices and users. Thus, all contributors to society, whether young or old members, can use these mobile apps. The use of these apps eases our daily lives, and all customers who need any type of service can access it easily, comfortably, and efficiently through mobile apps. Particularly, Saudi Arabia greatly depends on digital services to assist people and visitors. Such mobile devices are used in organizing daily work schedules and services, particularly during two large occasions, Umrah and Hajj. However, pilgrims encounter mobile app issues such as slowness, conflict, unreliability, or user-unfriendliness. Pilgrims comment on these issues on mobile app platforms through reviews of their experiences with these digital services. Scholars have made several attempts to solve such mobile issues by reporting bugs or non-functional requirements by utilizing user comments. However, solving such issues is a great challenge, and the issues still exist. Therefore, this study aims to propose a hybrid deep learning model to classify and predict mobile app software issues encountered by millions of pilgrims during the Hajj and Umrah periods from the user perspective. Firstly, a dataset was constructed using user-generated comments from relevant mobile apps using natural language processing methods, including information extraction, the annotation process, and pre-processing steps, considering a multi-class classification problem. Then, several experiments were conducted using common machine learning classifiers, Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM), and Convolutional Neural Network Long Short-Term Memory (CNN-LSTM) architectures, to examine the performance of the proposed model. Results show 96% in F1-score and accuracy, and the proposed model outperformed the mentioned models.

KEYWORDS

Mobile apps issues; play store; user comments; deep learning; LSTM; bidirectional LSTM

1 Introduction

Mobile application development and usage have grown, particularly during the COVID-19 pandemic owing to the increased usage of mobile apps in daily activities. Several mobile apps have been developed to assist with food delivery, grocery shopping, online shopping, appointment-making, health services, and other services that help people. These mobile apps are uploaded to marketplaces such as Google Play Store, Apple's App Store, Samsung Galaxy Store, Amazon App Store, and the

Apple App Store to be available to the public [1]. These marketplaces allow users to express their experiences and suggestions and provide feedback through written reviews and ratings of mobile apps, which are considered indicators of mobile app quality [2]. Users read the review comments before downloading such mobile apps. Such information about mobile apps can be an indicator of their weaknesses and usability issues before downloading and using the apps. Such information also helps users to obtain knowledge about mobile apps. Moreover, users express their feelings when downloading mobile apps by giving comments. Furthermore, user reviews contain information with technical value to the developer as a feedback channel. Developers can use user reviews as a source to improve the software functionality of mobile apps. Based on such information, mobile app developers can improve software usability and user satisfaction by analyzing user reviews.

Scholars have attempted to mine user reviews of mobile apps through many methods, such as bug reports and sentiment analysis, using the rating mechanism from 0 to 5, which is considered a quality measurement for mobile apps. The user comments cover a wide range of topics and issues, such as performance, battery issues, security, graphical user interface, registration, and updates. Therefore, the issues of mobile apps can be categorized and summarized using these methods. For the same purpose, some researchers paid attention to mining non-functional requirements (NRFs) as measures of the quality of mobile apps. In this way, researchers extract information and classify mobile app issues from the users' end perspective. NFR can be seen from different perspectives that are important to evaluate mobile apps, such as the usability of a mobile app's performance when installing, opening, and meeting the tasks or functions [3]. In addition, reliability measures the consistency of the mobile app functions [4]. Moreover, existing methods that have been utilized in mining mobile app issues include classifications [1,2,5–7], summarization [8–11], comparing features amongst mobile apps [12–14], sentiment analysis [15–18], and mining wearable devices [19–22].

Generally, users face difficulties in using mobile apps for several reasons. These issues can be realized from the starting point of downloading the mobile apps, registration process, graphical user interface, or use of the functions of mobile apps. As mentioned earlier, scholars have proposed suggestions, methods, approaches, and tools for addressing such issues. In traditional methods, qualitative approaches and surveys are conducted to gather user feedback and performance data analysis, whereas in recent methods, machine learning (ML), deep learning (DL), and natural language processing approaches are employed. The majority of these methods are based on the English language and focus on mining information from different perspectives and purposes. However, to the best of our knowledge, this case is an excellent challenge in Arabic society, where there is no labeled dataset and no prominence given to Hajj and Umrah services. Furthermore, an issue exists, and users are not satisfied.

Therefore, this study aims to propose a combined two Bidirectional Long Short-Term Memory (BiLSTM) model that can detect and classify mobile app issues related to Hajj and Umrah services through user-generated comments. Two BiLSTM layers can be used to better understand context. Simple patterns are detected by the first layer, and they are combined by the second layer to understand more intricate meanings. Comparable to having two sets of eyes on the text, this increases the model's capacity for learning and accuracy. When it comes to Hierarchical feature learning, the BiLSTM layers play an important by stacking the layers, whereby lower-level features are extracted by the first layer and more abstract representations are captured by the second layer. By enhancing the representational capability of the model, addressing the vanishing gradient issue, and producing ensemble-like effects, this method improves generalization and performance on challenging text classification tasks. The number of hidden units and layers, which determines the network's capacity; the dropout rate, which aids in preventing overfitting; the learning rate, which influences the optimization process; batch size,

which affects training stability; activation functions, which affect information propagation; weight initialization methods, which ensure proper convergence; and the optimizer, which determines how the model updates its weight.

In addition, a novel dataset is introduced based on information collected from user-generated comments on the most commonly used mobile apps serving during the Umrah and Hajj periods. The novel dataset was constructed from two languages (i.e., Arabic and English) using several natural language processing methods, such as in-formation extraction, pre-processing steps, and annotation process. A part of this dataset was translated from English to Modern Arabic standards using Google Translation API and also included dialects from different Arabic countries, such as Egypt, Saudi Arabia, Jordan, Yemen, and the Gulf area. The proposed dataset was considered imbalanced, which consists of 23,559 user comments, and was composed of five classes, namely, not working, registration issues, interface issues, update issues, and login issues after the filiation process of 76,351 user comments. Several experiments have been conducted to examine the performance of the proposed model. In these experiments, ML classifiers used, namely, support vector machines (SVM), Decision Tree (DT), Support Vector Classifier (SVC), and Logistic Regression (LR) and DL architecture, are ANN, LSTM, and CNN-LSTM. The ML classifiers and DL architecture have been examined based on the multiple hyper-parameters. The experimental results show that the proposed model outperforms the mentioned classifies in ML and DL in terms of score and accuracy. The model performance recorded 96% in terms of score. The main contributions of this study can be summarized as follows:

- Develop a hybrid BiLSTM model based on fusion to detect and classify Hajj and Umrah mobile app issues that can help developers improve mobile apps.
- Construct a bilingual dataset utilizing user-generated comments from Hajj and Umrah mobile apps.
- Apply different MCLs and DLs in several experiments on the proposed dataset to measure the performance of the proposed model.
- Examine the different weight methods that are calculated in the concatenate layer of two BiLSTM models.

The remainder part of this paper is organized as follows: Section 2 presents the related studies. The methods that are used to carry out this study are described in Section 3. Section 4 explains the results and discussion, and the conducting of this paper is presented in Section 5.

2 Related Studies

This section describes the related studies in user text review classification and prior research that concentrates on mining user comments from the perspective of improving the functionality of mobile apps. In addition, the most common apps that are used in the Hajj and Umrah. Therefore, Vermetten et al. [23] studied and evaluated the usability issues older users had with two mHealth apps and utilized MOLD-US, a recently developed framework for aging barriers, as a categorization method to pinpoint the root of these issues using the Think Aloud Protocol. Design principles should take into account older persons' declining cognitive abilities, physical limitations, and motivational impediments. Existing expertise in building interfaces for older target groups is not properly applied. It was found that 31% of these issues in the selected apps were categorized as cognitive issues.

Similarly, Kumar et al. [24] discussed ten heuristics that were given by Nielsen [25] for the heuristic evaluation of mobile learning applications, but due to their generic character, these heuristics are not very useful. This study analyzed 16 usability tests from the literature using new heuristics that

were established to augment Nielsen's heuristics. The validation results show that while evaluating mobile learning applications, evaluators can identify more usability issues by using newly created heuristics. Delikostidis et al. [26] investigated the usability of mobile applications within virtual environments considering several criteria, including the effectiveness and efficiency of task completion, users' satisfaction, and the number of errors users make. Empirical research was carried out by Moumane et al. [27] using a set of measures to assess the usability of mobile applications that are run on various mobile operating systems. Formerly, Flora et al. [28] investigated the most important characteristics that define mobile applications to facilitate the delivery of valuable user-friendly mobile apps to meet users' requirements. Méndez Porras et al. [29] performed a systematic literature review to identify and collect required evidence regarding automated testing of mobile applications. The usability measurement is done by taking into account three factors, i.e., effectiveness, efficiency, and satisfaction. In this process, some other factors, such as cognitive load, may be overlooked. Harrison et al. [30] designed a novel usability model, namely, people at the center of mobile application development. In addition, they reviewed mobile applications for specific fields.

In addition, Ebrahimi et al. [10] stated that the proliferation of mobile applications (apps) over the past decade has imposed unprecedented challenges on end-user privacy. Collection tactics have raised major privacy concerns among mobile app users. 2.6 million app reviews were sampled from three different application domains. Similarly, di Sorbo et al. [31] developed and have come up with a creative solution called the SURF (Summarizer of User Reviews Feedback) to shorten the time needed for user feedback analysis. 17 mobile apps from 23 developers and academics were subjected to an end-to-end evaluation of SURF; the findings revealed good accuracy in summarizing reviews and the value of suggested adjustments. SURF aids developers in comprehending user needs better than manual user request analysis and future program change planning. 68.75% of users said that it was easy to read and understand; however, 12.50% reviewed SURF and concluded it to be hard to read and understand; and then there are 18.75% who said it is somewhat readable and understandable. Jha et al. [32] used a unique method for app review mining from the app store and Twitter. The method proposed in this paper is based on frame semantics. The investigation, which makes use of three datasets of app store evaluations, demonstrates how semantic frames can facilitate an effective classification procedure for reviews. A review categorization and summarization package called MARC 2.0 uses the techniques that were looked at in the investigation. It was also found that 50% of the collected tweets contained maintenance requests.

Jha et al. [2] researched a two-phase study that mines NFRs from user evaluations that are found in app stores for mobile applications. A qualitative study of 6,000 user evaluations from a variety of iOS app categories was done in the first stage utilizing the dataset. The findings revealed that users in various app categories tended to raise various forms of NFRs, with 40% of the reviews in the sample indicating at least one sort of NFR. with an accuracy precision of 70% (range – 80%) and an average recall of 86% as a result. However, Lu et al. [33] studied and combined three machine learning algorithms—Naive Bayes, J48, and Wrapping four classification techniques—BoW, TF-IDF, CHI2, and AUR-BoW—to classify user reviews into four categories: NFRs (reliability, usability, portability, and performance), Functional Requirements (FRs), and Others. The results demonstrated that adding user evaluations can improve classification outcomes. Hasan et al. [34] analyzed 4.5 million Google reviews alongside 126,686 responses on 2,328 top free apps on the Google Play Store. to show and demonstrate how reviews are dynamic and how users and developers use the response mechanism as a basic user assistance tool. Owners of app stores might improve the response process by automatically flagging reviews that likely need a response. Chen et al. [1] extracted the most insightful user reviews, automatically grouping them using topic modeling, prioritizing them using an effective

review ranking scheme, and presenting the groups of the most insightful reviews via an intuitive visualization approach, AR-Miner is a novel computational framework for app review mining. The apps used were taken from the Google Play Store alongside the comments, which were picked out randomly for the proposed apps.

Noei et al. [35] studied the importance of high star ratings in mobile app markets. This study showed that there are several main themes in different categories and each category has a certain set of main themes. Several scholars have designed different mobile applications applicable to various services. Several master's theses have been carried out on creating mobile applications that can guide Malaysian pilgrims [36–42]. In a master's thesis [41], augmented reality was proposed to guide pilgrims. A new framework was suggested by [43], which serves as a crowdsourcing platform for defining the pilgrims' needs. In addition, reference [44] designed a mobile application serving as a dictionary for pilgrims who speak the Malay language. It has the capacity for translating three languages, i.e., Malay, Arabic, and English, to each other. The authors in [45] suggested the utilization of a mobile phone for tracking purposes. This strategy could aid the guides (Mutawwif) of pilgrimage groups in recognizing their movements and determining their locations. A dynamic signage system was designed by [46], which can be used as a mobile application that informs pilgrims regarding the crowd status around the Holy Ka'bah. Pilgrims can also use this application to automatically count the number of Tawaf rounds performed. Another mobile application was designed by [47], which is capable of guiding pilgrims when doing Hajj and Umrah. It can also serve as a Tawaf counter and location tracker.

Zhou et al. [48], introduced a study on BiLSTM models based on attention for recognizing personality from user-generated content. To find out how emoji information affects personality recognition task performance, the study suggests two innovative attention-based BiLSTM architectures that combine textual and emoji data. In addition, Zhang et al. [49] proposed a sentiment analysis and intuitionistic fuzzy TODIM method-based model for product selection. The suggested method seeks to use information from online reviews to help consumers decide what to buy. Similarly, Zhou et al. [50], highlighted the problems posed by the proliferation of online reviews, which has resulted in information overload. The study suggests a brand-new feature ranking method to find and extract significant features from online product reviews.

The author in [51] introduced a service for identifying an exact location in Hajj. Another research [52] offered a mobile translation application for the facilitation of communication among pilgrims. The "Al-Hajj" is a mobile application introduced by [53] to help pilgrims understand Hajj-related deeds. It can exhibit four interactive features, i.e., a map, checklist, motivational messages, and contact list. Reference [54] investigated how smartphone apps could be utilized to conduct surveys. Analytical research was carried out by [55] into the mobile applications related to the Hajj practices, which are offered on Google Play. The applications identified in this regard were analyzed by taking into account the following criteria: Supported languages, offered services, installation rates, and repetition of the services and names of the applications.

3 Materials and Methods

The methods that are used to carry out the proposed model are explained in this section. These methods are user comment extraction, data annotation process, data pre-processing, model architectures, and model evaluation methods. Fig. 1 shows an overview of the approach used in this research methodology.

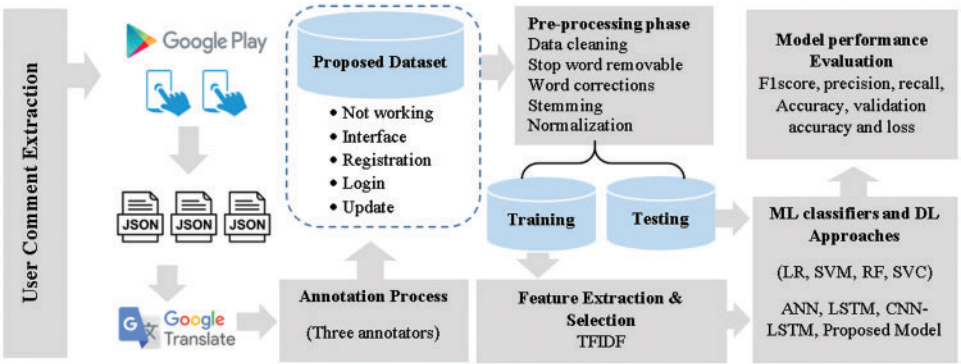


Figure 1: Research methodology approach

3.1 User Comment Extraction

The first task is to extract the information from user-generated comments on the mobile apps from the Google Play Store to detect the issues of the mobile apps related to Hajj and Umrah services. The main criteria used in selecting and extracting the data are identified mobile apps that are used in Hajj and Umrah focus on user comments, and all the textual data is in Arabic and English, with different dialects, such as in Saudi, Egypt, Jordan, Gulf Area, and Yemen. The mobile apps were selected based on the most popular applications used by the pilgrims and the ones that provide the most services to them. Then, the seven mobile apps were selected, and the user comment extraction processes were executed.

In the extraction process, the Google Store API using the 'google_play_scraper' Python library has been used with Python 3.7 programming language using Google Colab. A total of 76,351 user-generated comments were extracted from mobile apps that provide services to pilgrims during the period January 2020 to July 2023. The user comments are extracted from the Google Play Store in the form of a comma-separated values (CSV) file that contains several attributes. Most of the attributes that were not relevant to the objective of this study were removed, and only the user comments attribute remained. The data annotation process utilizes this CSV in its second step. In addition, in this stage, data cleaning processes have occurred, such as removing noise, special characters, numbers, and emojis.

3.2 Data Annotation

In this step, the data annotation process begins, and the classes are identified to give labels before the classes are identified, and the labels are provided. During this stage, three experts and Ph.D. holders in the area of software engineering read comments randomly to obtain the initial idea of the user comments and list the issues. Initially, the number of classes was identified by 12 labels, then reduced to five classes only. This process is time-consuming. The experts spent much time classifying user-generated comments after a careful reading process to assign the comments to the relevant class. The agreement to assign user comments to the class should be obtained with at least the approval of two annotators; otherwise, the user comment is excluded. In addition, Cohen's kappa measure was used which showed the agreement between the annotators 82% which indicates an almost perfect agreement. Table 1 shows the proposed dataset, and Fig. 2 shows the general view of the dataset in the form of the word cloud.

Table 1: Dataset description

No.	Class	No. of comments
01	Not working	7,055
02	Registration issues	3,668
03	Interface issues	2,014
04	Update issues	6,165
05	Login issues	4,657
Total		**23,559**

Figure 2: Word cloud

The five classes have been identified based on their content and double-checked for each user comment by experts. Table 2 presents a specific definition for each class. The user comment was found to be written in Arabic, English, or Arabizi. The English comments were translated into the modern Arabic language. In the Arabic comments, the comments come in a different dialect, as the Arabic language can contain many dialects of the Arabic world. An example of the user comments are shown in Table 3. And Table 4 shows an example for each class.

Table 2: Classes and its description

No.	Class	Class code	Definition
01	Not working	1	They system is hanging or not open
02	Registration issues	2	There is issues during registrations such as request for resident number while some people are visitors
03	Interface issues	3	This is happened with difficult of graphical user interface issues
04	Update issues	4	This when the system ask user for updates
05	Login issues	5	Matters that related to login process

Table 3: Example of the user comments and classes

Class code	Example	Translated to English
1	التطبيق يفصل من تلقاء نفسه عند الدخول الي حسابي برقم الإقامة	The application close by itself when entering my account with the residence number
2	يطلب مني رقم التأشيرة. وانا اريد استخراج تصريح للعمرة اصلاً لو كان عندي تصريح ماكنت احتجت للتطبيق	Application asks me for the visa number. And I want to obtain a permit for Umrah in the first place. If I had a permit, I would not need to apply
3	التطبيق غير مكتمل سيء لا يظهر قيمة السلعة في العربة وايضا بعد ان تستوضح معلومات سلعة لن يرجعك الي مكان السلعة في القائمة التي كنت تختار منها بل تعود للرئيسية وتضطر الى البحث من جديد في القوائم بالرئيسية اضافة الى خروجه من الخدمة لمرات عديدة عندما تحاول اختيار سلع ولو ينقلب عليك الجوال تعيد من الصفر. نامل تطوير التطبيق	The application is incomplete and bad. It does not show the value of the commodity in the cart. Also, after you clarify the information of the commodity, it will not return you to the place of the commodity in the list from which you were choosing. Rather, you return to the main one and have to search again in the main menus, in addition to leaving the service many times when you try to choose goods, even if Mobile turns on you restart from scratch. We hope to develop the application
4	يعطيني تحديث ولا يحدث وحذفت التطبيق وحملته ونفس المشكلة	Application gives me an update and it does not happen and I deleted the application and downloaded it and the same problem
5	يطلع اسم المستخدم وكلمة المرور غير صحيحه رغم ان كل شي صحيح ورسمي افيدوناجزيتم خيراً	The username and password are incorrect, although everything is correct and official

Table 4: ML classifiers and hyper parameters

No.	Classifier	Hyper parameters
01	LR	C = 1.0, intercept_scaling = 1, max_iter = 1000, penalty = 'l2', solver = 'lbfgs', tol = 0.0001, verbose = 0
02	SVM	C = 10, gamma = 0.001, kernel = 'rbf'
03	SVC	C = 15, gamma = 0.01, kernel = 'rbf'
04	RF	Criterion: 'gini', max_depth: 150, max_features: 'auto'. min_samples_leaf: 4, min_samples_split: 7, n_estimators': 150

3.3 Data Pre-Processing

In this step, the input of ML, DL and the proposed model used is the proposed dataset. Several steps are required for preparation for user-generated comment inputs. The pre-processing steps are

performed to obtain high accuracy by training the models to predict the classes for user comments. To prepare the user comments for the next phase, several pre-processing steps were employed. These steps remove noise from the user comments, such as special characters or emojis, remove numerical data, remove stop words and normalization for the Arabic language, use a regular expression to deal with white spaces and use correction methods for Arabic words and stemming steps. Then, the user comments were translated to TFIDF word representation as presented in mathematical Eq. (1), and the bigrams were utilized with 2,500 features that achieved the best accuracy.

$$W_{ij} = TF_{ij}X \log\left(\frac{Total\ user\ comments}{Number\ of\ user\ comments\ conatin\ word\ j}\right) \tag{1}$$

whereas W denote the weight of word "i" in user comment "j" and TF is the term frequency of the word "i".

3.4 Building Models

This section describes the most commonly used methods that are used in textual classification problems. Scholars selected ML classifiers and DL architectures that are used to evaluate the performance of the proposed model. ML classifiers are SVM, DT, LR and SVC. SVM is the linear classification algorithm that uses a line or a hyperplane to separate data points in dimension space (user comments) into classes by breaking the multi-classification problem into sub-problems to deal with them as a binary classification problem. In the same way as SVM, which is a variant of SVM, the SVC algorithm works by finding the best hyperplane to classify data points. Meanwhile, LR is a statistical algorithm used for predicting classification based on categorical variables utilising logic functions. LR is based on a modified function of linear regression that is used for continuous prediction. DT represents the problem in tree structures that consist of nodes in the form of children and parents and is then computed by assigning user comments to the classes. This way is based on the concept of probability that uses entropy and information gain.

ANN is a simple architecture of DL. ANN consists of interconnected neurons in layer structures, which are input, hidden, and output layers, as shown in Fig. 3. In the input layer, the input is user-generated comments represented in the form of TF-IDF. Meanwhile, in the hidden layers, which can consist of many layers, the input is used to train the ANN to classify the comments into five defined classes in the output layer. The ANN works based on the mathematical formula in Eq. (2). Based on the calculation of this formula, the activation function is used to decide whether the output of the neuron is activated or not.

$$Y = W_1X_1 + W_1X_1 + W_nX_{n1} + b \tag{2}$$

where X is the input for the ANN, which is TF-IDF representing the term (word) of the user comment. Therefore, several terms are represented in X_n and n is the number of the TFIDF, whereas W is a weight that is assigned to each neuron in the hidden layers, and b is a small value known as bais. Then, Y is considered the input for the next neuron.

The LSTM architecture is the type of recurrent neural network that is used to learn based on sequence data. LSTM knows a sequence processing model. In addition, LSTM stores terms for long-term dependencies, which are the terms in the user comments. Therefore, LSTM remembers the previous terms required in textual data. Thus, this structure of DL is utilized in natural language processing as it improves in identifying/predicting the class of user comments to which classes belong. Fig. 4 shows the LSTM structure.

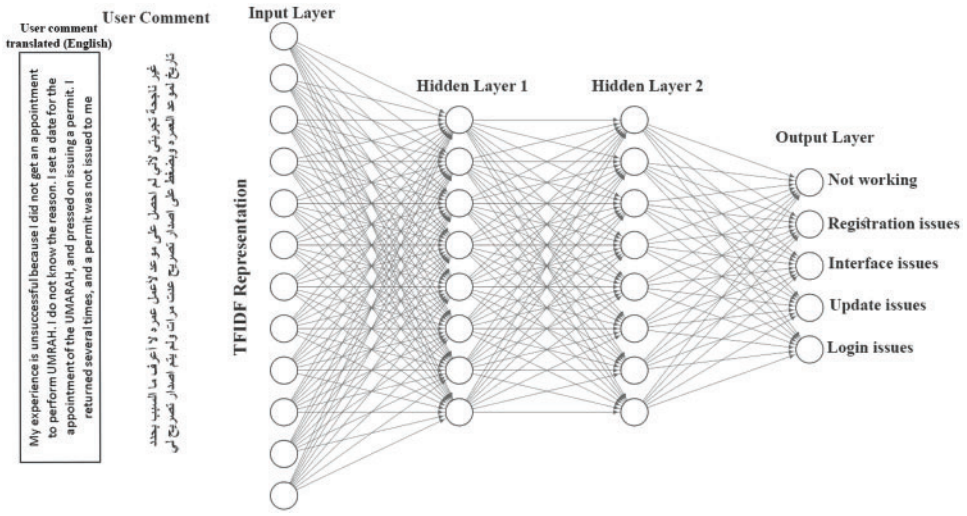


Figure 3: ANN architecture

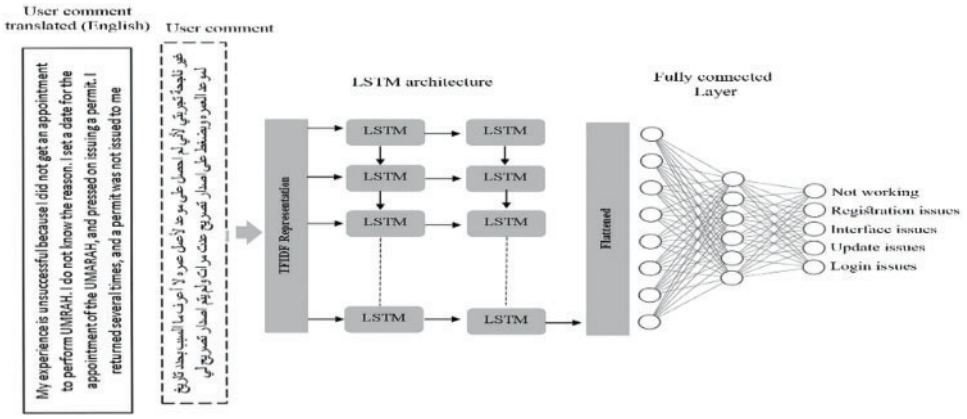


Figure 4: LSTM architecture

The CNN-LSTM consists of two consecutive stages: CNN architecture followed by LSTM. CNN is utilized for extracting features (terms) from the input data. Meanwhile, LSTM used that data to train the neural network on the user comments to predict the classes of mobile app issues based on user-generated comments. Fig. 5 depicts the CNN-LSTM architecture.

3.5 Proposed Model

The proposed model is the combination of two BiLSTM models. BiLSTM architecture is built based on the LSTM model architecture. BiLSTM has an added reversed way of information, where the network works in both directions, forward and backwards, to overcome the weakness of LSTM, that is, remembering the previous only, which is the time of t only and the future context of information (forward propagation). Therefore, BiLSTM consists of two LSTM. This model is used mainly for natural language processing and is a powerful way to deal with dependencies between words/terms in the user comments. In this way, the network learns in a bidirectional way between time steps/sequences of data based on the dependency concept. Generally, BiLSTM capture bidirectional

context and handle long-term dependencies in sequential data. By this way, BiLSTM allows to capture contextual information from both past and future tokens. Therefore, it provides a more comprehensive understanding of the text. Fig. 6 shows the model architecture.

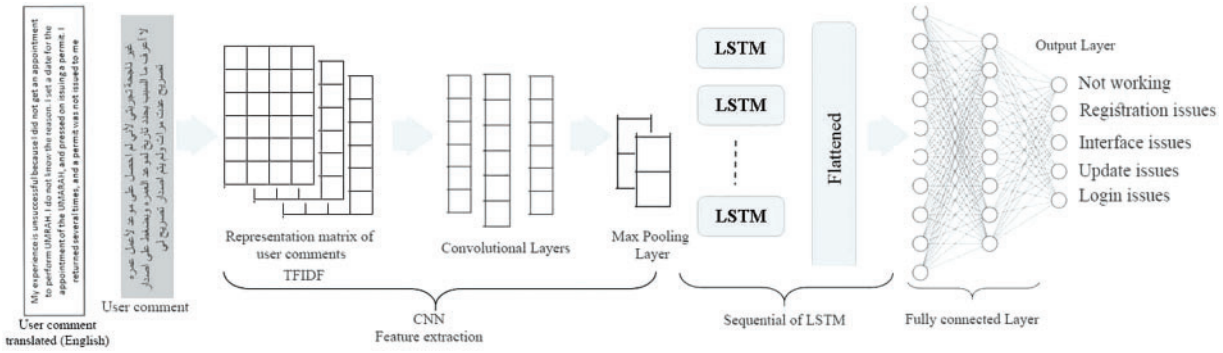


Figure 5: CNN-LSTM architecture

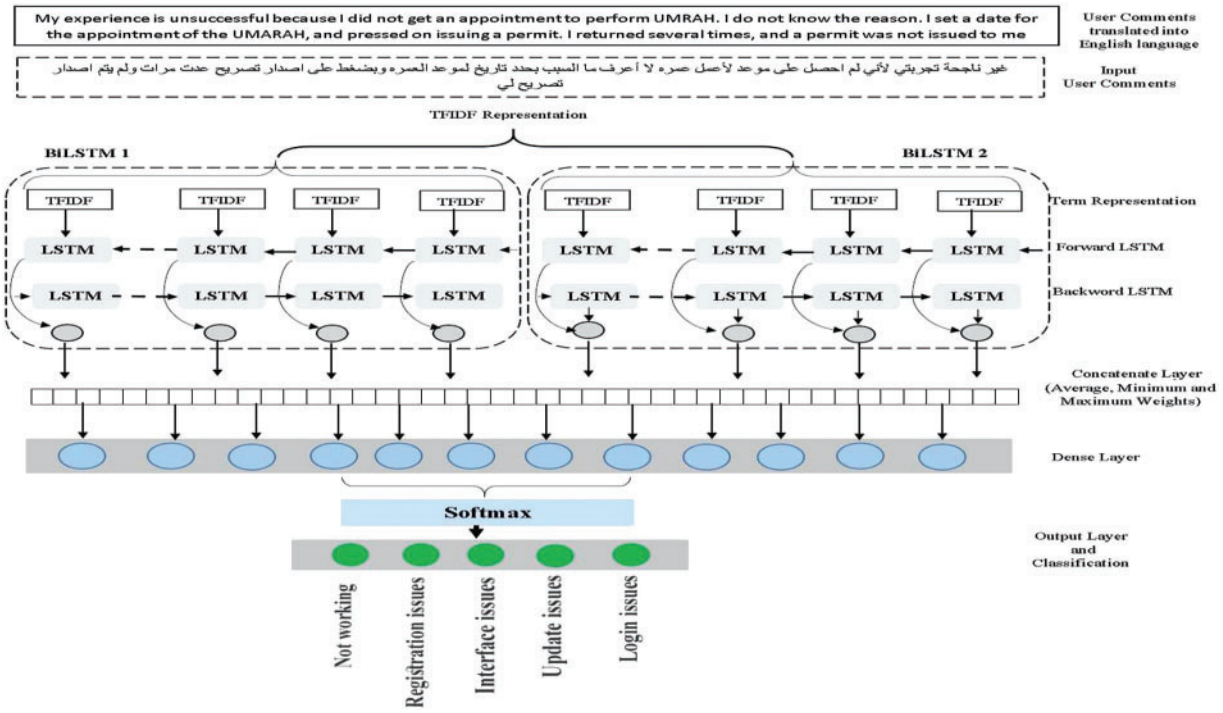


Figure 6: Proposed combined BiLSTM model

Fig. 6 illustrates the architecture of the proposed model that consists of two BiLSTM architectures. This architecture is composed of five phases, namely, input, TFIDF representation, two BiLSTM architectures, combined layer, and output layer. In the input phase, user-generated comments about the mobile app's issues are received, and then the user-generated comment about the mobile app's issues is represented in TF-IDF. These inputs are fed to two LSTM models working in a bidirectional way. The results of the two models are combined in different ways: Average, minimum and maximum weight. Then, the results of the weight are processed to the dense layer, which is ANN. The final decision

for the classification of the user comments about the mobile apps issues is classified into one of the five classes using the Softmax function as represented in mathematical Eq. (3), which is known as the control of information flow and regulation of the values in the gate. In the softmax activation function, $\overrightarrow{Z}$ is the input vector (list of words), K is the number of classes (multi-classes) and e^{zj} is the exponential function for the output vector, whereas the Tanh activation function is used in hidden layers, which can be represented in mathematical Eq. (4):

$$Softmax\left(\overrightarrow{z}\right) = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \tag{3}$$

$$Tanh\,(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \tag{4}$$

BiLSTM calculates the input of the user-generated comments in sequence order. The input of unique words, denoted by x, can be shown in the mathematical formula in Eq. (5) as follows:

$$x = (x_1, x_2, x_3, \ldots x_n) \tag{5}$$

These inputs pass through in a forward direction from left to right, as represented in the mathematical formula in Eq. (6) as follows:

$$h_t = (\overrightarrow{h_1}, \overrightarrow{h_2}, \overrightarrow{h_3} \ldots\ldots\ldots, \overrightarrow{h_n}) \tag{6}$$

and the second LSTM in a backwards direction from left to right, as represented in the mathematical formula in Eq. (7).

$$h_t = (\overleftarrow{h_1}, \overleftarrow{h_2}, \overleftarrow{h_3} \ldots\ldots\ldots, \overleftarrow{h_n}) \tag{7}$$

and concatenation of forward and backwards can be summarised in the following mathematical formula in Eqs. (8) and (9), respectively, and the concatenation of forward and backwards can be stated in the mathematical formula in Eq. (10).

$$h_t = LSTM(x_t, \overrightarrow{h_{t-1}}) \tag{8}$$

$$h_t = LSTM(x_t, \overleftarrow{h_{t+1}}) \tag{9}$$

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h}_t] \tag{10}$$

whereas the output is y, as represented by the mathematical formula in Eq. (11).

$$y = (y_1, y_2, y_3, \ldots y_n) \tag{11}$$

Details of the procedures that are use to detect and classify the mobile app issues related to Hajj and Umrah is presented in Algorithm 1.

Algorithm 1: Detecting and classifying mobile apps issues related to Hajj and Umarah

0 **INPUT:**
1 User generated comments about Hajj and Umarah related issues;
2 UC denote user generated comments;
3 UC represented by TFIDF;
4 Total number of UC = N;

(Continued)

Algorithm 1 (continued)

```
C denote predefine classes;
DL denote deep learning model;
OUTPUT:  UC unlabeled dataset
BEGIN
While  N=0 DO                                   # for training phase
  For all  UC→data_preprocessing_phase;
    data_preprocessing_phase→proceesed_UC;
    Processed_UC → TIFDF;
     TFIDF_All_UC ← TFIDF;
     DL←TFIDF_All_UC;
        If  TFIDF for each UC in C then
            C←UC;
        End If
      END;
   For each input UC                            # testing phase or actual comment
     TFIDF_UC;
    DL←TFIDF_UC;
        If  TFIDF for each UC in C then
            C←UC;
       End If;
     DL →accuracy measures;
END;
```

3.6 Evaluation Methods

The most commonly used measures have been utilized to evaluate the performance of the proposed model used to classify and predicate the mobile app issues related to Hajj and Umrah. Furthermore, the experimental results of the proposed model are compared with those of the existing and common classifiers of ML and DL approaches. These measures are recall, precision, F-score and accuracy, all of which are shown in the following mathematical Eqs. (12)–(15), respectively.

$$Precision = \frac{Retrieved\ Relevant\ User\ comments}{All\ Retrieved\ User\ Comments} \tag{12}$$

$$Recall = \frac{Retrieved\ Relevant\ User\ Comments}{All\ Relevent\ Comments} \tag{13}$$

$$F - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{14}$$

$$Accuracy = \frac{Numberof\ correct\ user\ comments}{Total\ number\ of\ user\ comments} \tag{15}$$

4 Results and Discussion

This section describes the experiment settings and hyperparameters used in the experiments. In addition, the experimental results of the ML classifiers, DL approaches and the proposed model are

discussed. These experiments were carried out to measure the performance of the proposed model to detect and predict Hajj and Umrah mobile app issues from user-generated reviews.

4.1 Experiment Settings

Several experiments were conducted using the most commonly used ML classifiers and DL architectures. These methods produced high accuracy, particularly with textual data, which indicates good model performance in classification problems. The classifiers in ML are LR [56,57], SVM [58,59], SVC [60,61] and RF, and those in DL architecture are ANN, LSTM and CNN-LSTM. The experimental results of these mentioned approaches are used to compare with the performance of the proposed model. All the experiments were executed using Python on Colab and Scikit-learn (Sklearn) library for ML classifiers, NLTK for pre-processing methods and Tensorflow and Keras for DL approaches using GPU. The novel introduced in the dataset described in Table 1 has been utilized for all experiments. After several experiments, the best tuning configurations for hyperparameters were identified, as shown in Table 4.

In ML classifier experiments, several experiments were conducted to train the aforementioned ML classifiers to detect the Hajj and Umrah mobile apps issue based on the proposed dataset. The best results are obtained using the aforementioned hyperparameter, as shown in Table 4. In addition, the validation and test accuracy are close to each other, and the difference is not more than 5%. Therefore, the overfitting or underfitting problem does not exist. Owing to the many records of the experiment results for the five classes in terms of precision, recall, F1-score and accuracy, the average of precision, recall and F1-score for the five classes of issues related to the Hajj and Umrah mobile app experiments is used, as presented in Table 5. In addition, Fig. 7 illustrates the precision, recall and F1-score for all classes using ML classifiers. The best model performance for the classifiers ever recorded is 92.74% and 92.16% using RF and SVC, respectively.

Table 5: Precision, recall, F1-score, and accuracy of ML classifiers

ML classifiers	precision	recall	F1-score	Accuracy
LR	85.00%	81.09%	82.71%	81.59%
RF	93.56%	91.64%	92.51%	92.74%
SVC	92.98%	90.57%	91.65%	92.16%
SVM	86.50%	83.52%	84.82%	85.09%

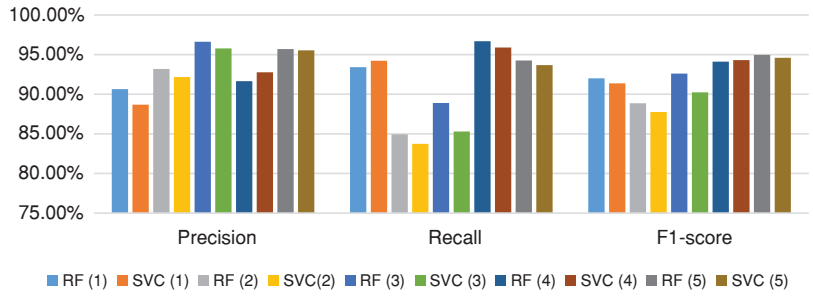


Figure 7: Comparison amongst precision, recall and F1-score for the five classes

In the DL experiments, the hyper-parameter and experiment setting was the same without any changes for all experiments of DL architecture, as presented in Table 6. All measurements for model performance used in the ML experiment are also utilized in the DL experiments.

Table 6: DL Hyperparameter

Parameters	ANN	LSTM	CNN-LSTM
Batch-size	128	32	64
Hidden layer-activation function	relu	Relu	Relu
Output layer-activation function	softmax	Softmax	Softmax
Dropout	0.5	0.5	0.5
Number of epochs	50	50	50
Loss-function	sparse_categorical_crossentropy		
Optimizer	Adam	Adam	Adam
Regularization	L2(0.01)	L2(0.01)	L2(0.01)

In the first experiment of DL using ANN architecture, the model produces 89% in terms of accuracy. In addition, the average recall is 86.60%, and the precision is 90.40. In this architecture, the sequential model using Keras on top of Tensorflow has been utilized. In the LSTM experiment, the model accuracy is the same as that produced by ANN experiment, with a slight improvement in terms of F1-score because the recall score is higher than the score in ANN. Overall, in both experiments, the accuracy and F1-score of the training and testing are close to each other. The model accuracy and F1-score decreased to 85% and 84.60%, respectively, where the decrease has been notable with the CNN-SLTM experiment. In addition, the training time is longer compared with the previous two experiments. Therefore, the worst model performance has been recorded using the CNN-LSTM architecture, which is 85%.

The BiLSTM model excels at text classification due to its unique ability to understand word context bidirectionally. As a result, using two BiLSTM models in text classification tasks can improve performance. It processes input sequences in both forward and reverse orientations, allowing each model to capture a more comprehensive context by taking preceding and subsequent words into account. Because of the bidirectional information flow, the models can better understand the relationships between words in a phrase, which is critical for successful classification. It is a sort of ensemble learning in which the characteristics of each model complement each other. This helps capture diverse parts of the data while boosting the model's overall resilience and generalization.

The experiment of the proposed model was carried out based on the hyperparamter as presented in Table 7. While the experiment results are shown in Table 8, which demonstrates the highest accuracy and F1-score produced using the proposed model. For the proposed model, precision and recall achieved the highest scores, which are 96% compared with the previous three experiments, in addition to the ML classifiers. Thus, the proposed model outperforms the ML classifiers and the ANN, LSTM, and CNN-LSTM in terms of accuracy and F1-score in detecting and predicting Hajj and Umrah mobile issues using five classes. The figure alongside Figs. 8–10 show the decreasing loss and the score in terms of accuracy for the ANN, LSTM, and CNN-LSTM models.

Table 7: Parameters used in the proposed model

Parameters	Description
Optimizer	Adam
Learning rate	0.01
Dense layer	5
Epochs	50
Batch size	32
BiLSTM nodes	10
Regularize	l2(0.001)
Loss function	categorical_crossentropy
Activation (internal)	relu
Activation (output)	Softmax
Drop out	50%

Table 8: Experiment results of DL approaches and proposed model

Architectures	Precision	Recall	F1-score	accuracy
ANN	90.40%	86.60%	88.20%	89.00%
LSTM	89.00%	87.80%	88.40%	89.00%
CNN-LSTM	85.00%	83.75%	84.60%	85.00%
Proposed model	**96.00%**	**96.00%**	**96.00%**	**96.00%**

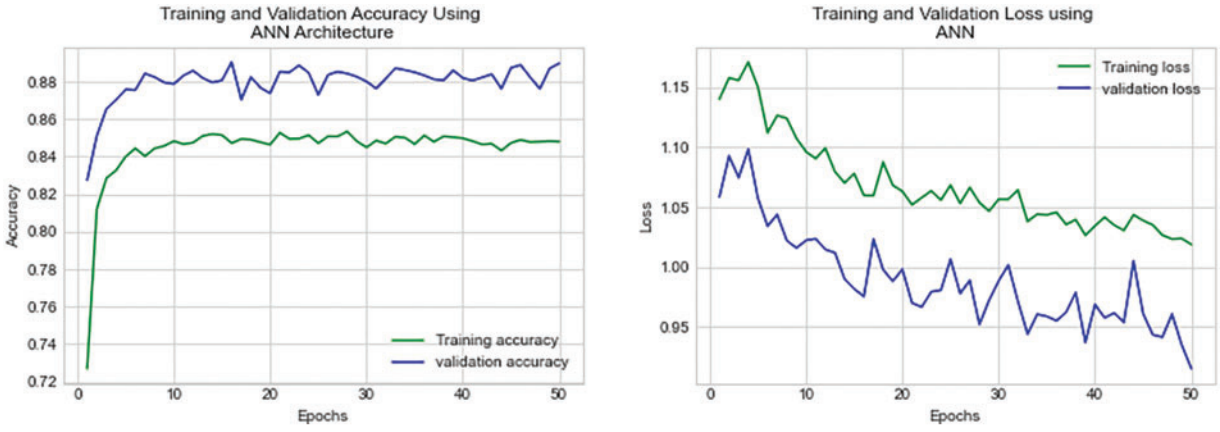


Figure 8: Training and testing accuracy and validation and loss using ANN architecture

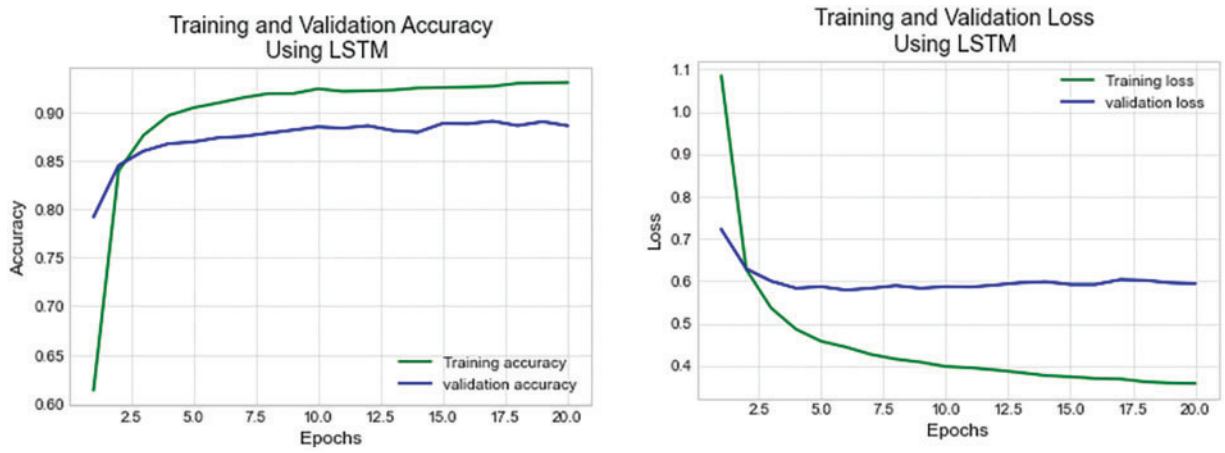


Figure 9: Training and testing accuracy and validation and loss using LSTM a architecture

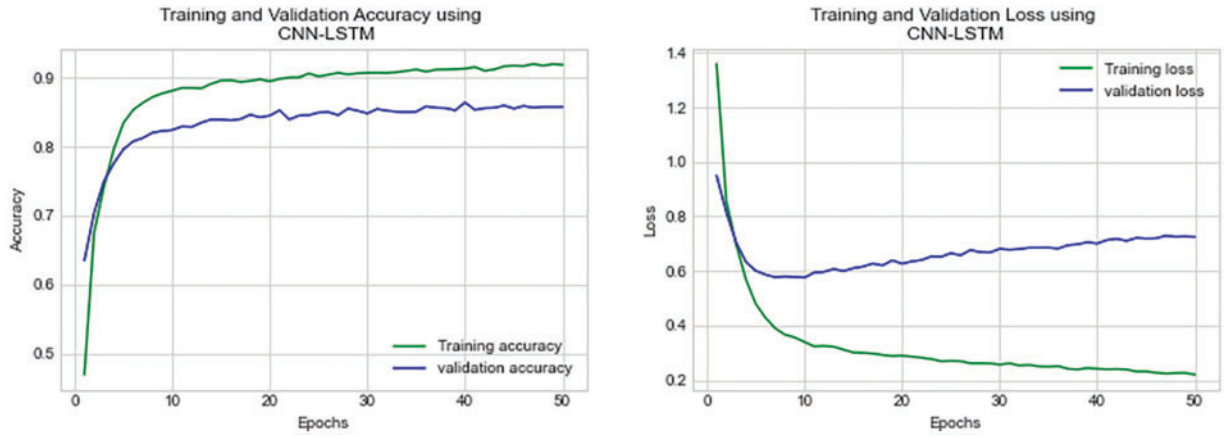
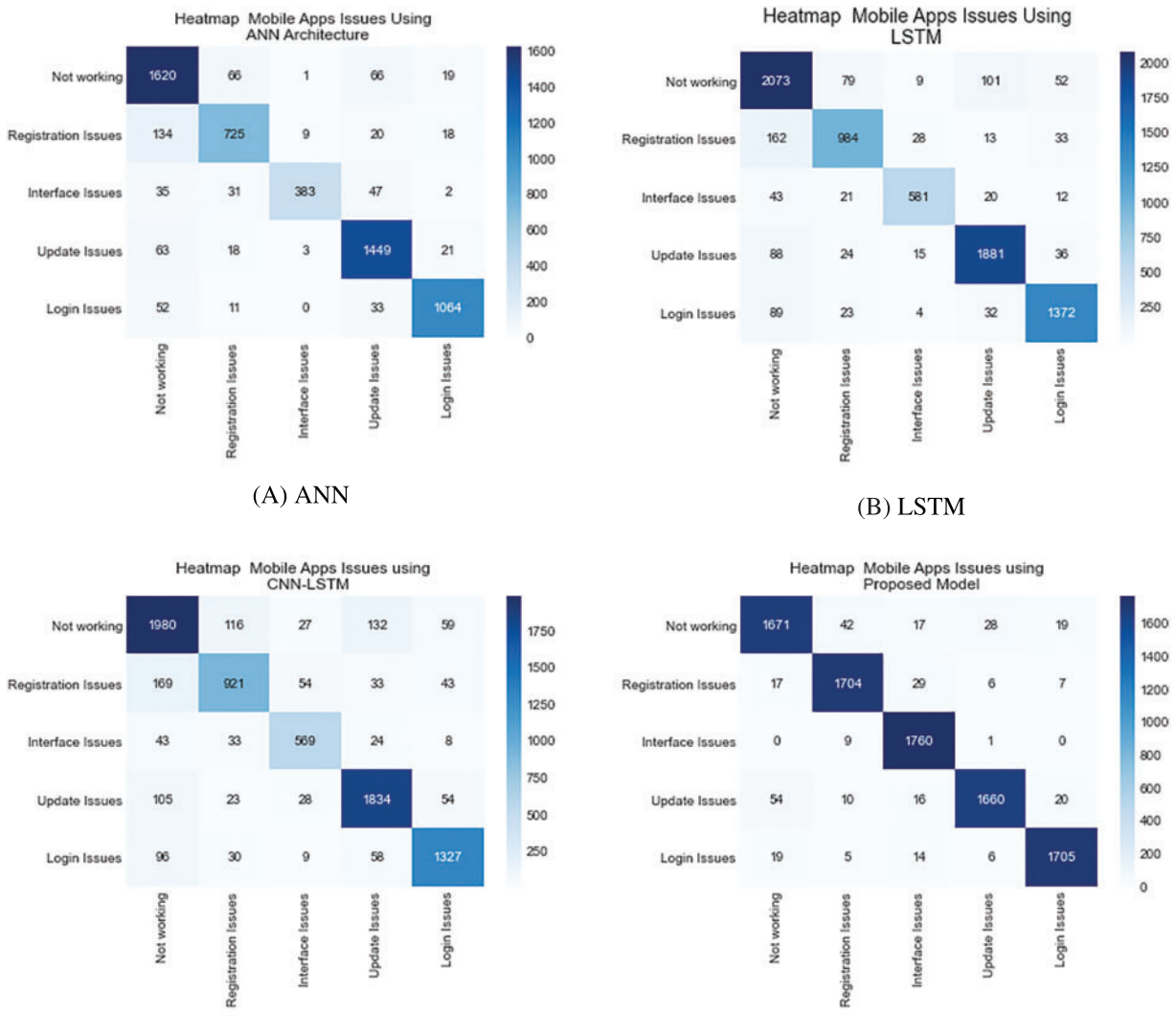


Figure 10: Training and testing accuracy and validation and loss using CNN-LSTM architecture

Fig. 11 illustrates the experiment results in the form of a confusion matrix for three DL architectures and the proposed model. In Table 8, it presents the performance of the proposed model compared with the deep learning models. The model performance in term of precision, recall, and F1-score of the proposed model for five classes is present in Table 9. While in Table 10, the model performance is compared with the state-of-the-art models that were proposed for text classification in different representations such as social media. Overall the proposed model has outperformed all the other models it was tested with in terms of accuracy. Which attained a score of 96%, which proves that processing input sequences in both forward and backward directions can help in understanding the relation between the words.

Fig. 12 shows the accuracy and validation of the proposed model, with the highest training and validation accuracies.



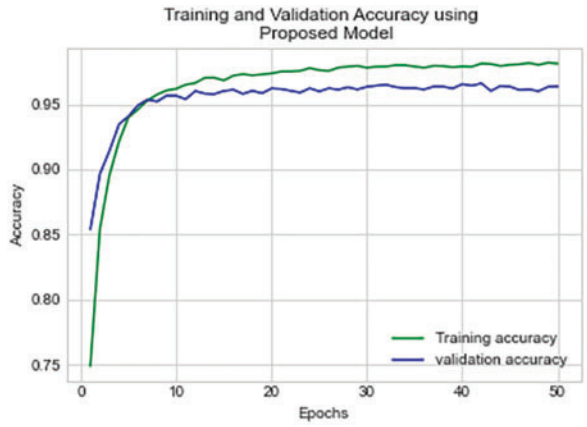
(A) ANN (B) LSTM

(C) CNN-LSTM (D) Proposed with concatenate weight

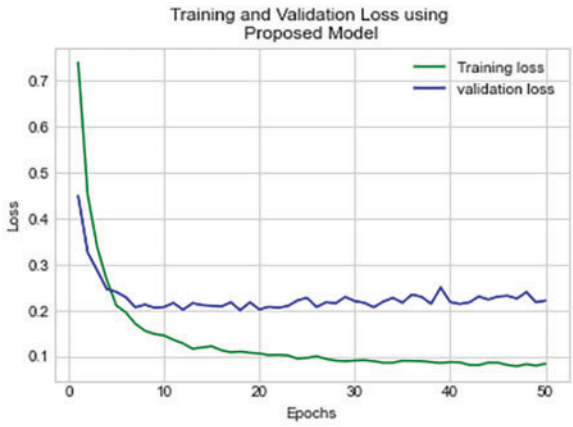
Figure 11: Confusion matrix of DL experiment

Table 9: precision, recall, and F1-score of proposed model for five classes

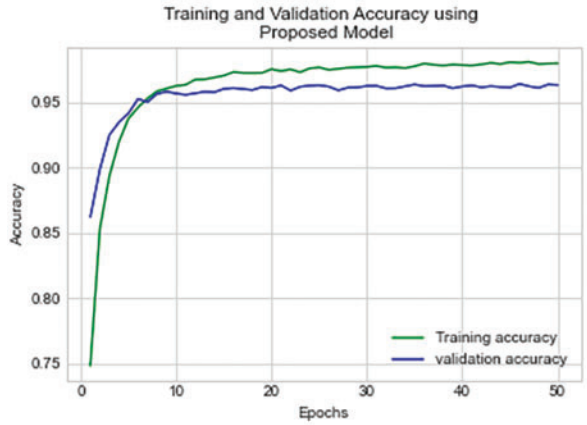
Class	Precision	Recall	F1-score
1	0.96%	0.93%	0.94%
2	0.96%	0.96%	0.96%
3	0.97%	0.98%	0.98%
4	0.97%	0.95%	0.96%
5	0.95%	0.98%	0.97%



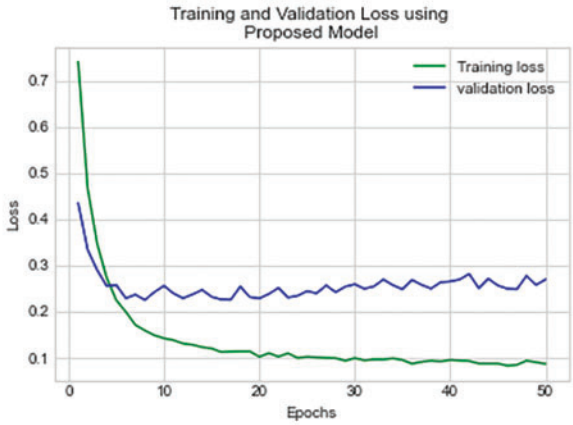
(A) Accuracy using concatenate weight method



(B) Loss using concatenate weight method



(C) Accuracy using Average weight method



(D) Loss using Average weight method

Figure 12: Accuracy, validation and loss of proposed model

Table 10: Compare accuracy of existing methods and propose model

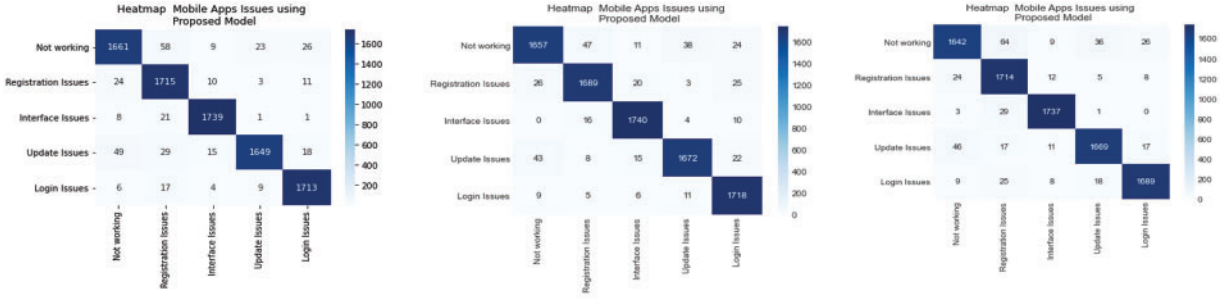
Authors	Domain	Accuracy	Dataset size	Language	Classes	Models
[62]	COVID-19	81.15%	563,079	English	Binary	LSTM
[63]	E-commerce	93.5%	100,000	Chinese	Binary	CNN+GRU
[64]	Shopping	93.1%	21,056 and 10,000	Chinese	Binary	GRU+attension
[65]	Social media	92%	197,566	English	Binary	XGBoost
[66]	Social media	95%	39,000	English	Binary	CNN
[67]	Social media	89%	100,000	English	Binary	BERT XLM
[68]	Mobile communications	82.14%	17,289	Turkish	Binary class	CNN+BiLSTM

(Continued)

Table 10 (continued)

Authors	Domain	Accuracy	Dataset size	Language	Classes	Models
[69]	Hospitality reviews	94%	1 Million	English	Multi-classes	LSTM
[48]	Personality traits	65%	9,917	Personality traits	Multi-classes	AttentionBiLSTM

As comprehended by Table 9, we can perceive that the proposed model had a solid 96% in all aspects of testing (precision, recall, F1-score, accuracy), which was much higher than all the other models compared with that between 83% and 90% amongst all the tested aspects. After examining Table 8, we can safely say that the proposed model was the best-performing model amongst all the other architectures as it out-scored all the other models with a significant accuracy of 96% in only 50 epochs that used the train proposed model based on the proposed dataset. Fig. 12 shows the accuracy and validation of the proposed model using the concatenate and the average weight in the combined layer. In addition, the maximum weight is used and has attained the same accuracy. Furthermore, the confusion matrix of the proposed model with different calculation methods in the combined layer are presented in Fig. 13. Table 10 demonstrates the performance of the proposed model in term of accuracy compared with the common existing models.



(A) Add weight (B) Average weight (C) Max weight

Figure 13: Confusion matrix of the proposed model with different calculation methods

5 Conclusions

This study proposes employing a fusion of two BiLSTM models to detect and classify mobile app issues that are related to Hajj and Umrah services. Millions of people use such applications and encounter issues, making customers unsatisfied. A novel dataset has been introduced based on a multi-class classification problem, which is extracted from user comments from the most common mobile apps using Arabic and English comments. The dataset consists of five classes identified by experts. The experimental results show that the performance of the proposed model outperforms the selected ML classifiers and DL architectures, and the score reached 96%. The model mode constructed based on two BiLSTM models achieved great results owing to its directional mechanism forward and backward, which are useful for sequence data, particularly textual data, which are the core of the classification problem. In these experiments, different types of combined calculation weights of features have been used, and no notable improvement has been recorded in terms of accuracy. There is

a slight exchangeable improvement in precision and recall. Future work can add more user-generated comments from many mobile apps and make it a bilingual classification, Arabic and English, by utilizing DL architecture and NLP methods to obtain higher accuracy, this can be achieved by adding a dataset that is larger and consists of user-generated comments from a varied range of mobile apps. Moreover, a bilingual classification approach including the languages of Arabic and English can be enhanced for the applicability of the model. Using advanced Deep Learning (DL architectures and Natural Language Processing (NLP) methods, such as multilingual embeddings or pre-trained language models, almost definitely leads to attaining a higher score in accuracy for classifying comments in both of the aforementioned languages. This development increases the understanding of the model through different domains and also addresses the diverse linguistic inherent in content that is user-generated.

Acknowledgement: The authors would like to express their gratitude for the valuable feedback and suggestions provided by all the anonymous reviewers and the editorial team.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: Conceptualization, Wael M. S. Yafooz and Abdullah Assaedi; Wael M. S. Yafooz and Abdullah Assaedi; Data curation, Wael M. S. Yafooz and Abdullah Assaedi; Formal analysis, Wael M. S. Yafooz; Investigation, Wael M. S. Yafooz; Methodology, Wael M. S. Yafooz and Abdullah Assaedi; Software, Abdullah Assaedi; Validation, Abdullah Assaedi; Visualization, Abdullah Assaedi; Writing–original draft, Wael M. S. Yafooz and Abdullah Assaedi. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The datasets can be accessed at https://www.kaggle.com/datasets/waelshaher/mobile-apps-issues.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] N. Chen, J. Lin, S. C. H. Hoi, X. Xiao, and B. Zhang, "AR-miner: Mining informative reviews for developers from mobile app marketplace," in *2014 36th Int. Con. on Eng.*, Hyderabad, India, May 31–Jun. 7, 2014, pp. 767–778. doi: 10.1145/2568225.2568263.

[2] N. Jha and A. Mahmoud, "Mining non-functional requirements from app store reviews," *Emp. Soft. Eng.*, pp. 3659–3695, Jun. 2019. doi: 10.1007/s10664-019-09716-7.

[3] N. M. Ismail, F. Ahmed, N. A. Kamaruddin, and R. Ibrahim, "A review on usability issues in mobile applications," *2016 IOSR J. Mob. Comput. App.*, vol. 3, no. 3, 2016. doi: 10.9790/0050-03034752.

[4] M. N. Islam, I. Islam, K. M. Munim, and A. N. Islam, "A review on the mobile applications developed for COVID-19: An exploratory analysis," *IEEE Access*, vol. 8, pp. 145601–145610, Aug. 2020. doi: 10.1109/ACCESS.2020.3015102.

[5] B. H. Leem and S. W. Eum, "Using text mining to measure mobile banking service quality," *Ind. Manage. Data Syst.*, vol. 5, pp. 993–1007, 2021. doi: 10.1108/IMDS-09-2020-0545.

[6] M. Hatamian, J. Serna, and K. Rannenberg, "Revealing the unrevealed: Mining smartphone users privacy perception on app markets," *Comput. Secur.*, vol. 83, pp. 332–353, 2019. doi: 10.1016/j.cose.2019.02.010.

[7] A. E. Yahya, A. Gharbi, W. M. S. Yafooz, and A. Al-Dhaqm, "A novel hybrid deep learning model for detecting and classifying non-functional requirements of mobile apps issues," *Electron.*, vol. 12, no. 5, pp. 1258, Mar. 2023. doi: 10.3390/electronics12051258.

[8] C. Tao, H. Guo, and Z. Huang, "Identifying security issues for mobile applications based on user review summarization," *Inform. Soft. Technol.*, vol. 122, no. 106290, pp. 1258, Jun. 2020. doi: 10.1016/j.infsof.2020.106290.
[9] X. Zhang *et al.*, "DSGPT: Domain-specific generative pre-training of transformers for text generation in e-commerce title and review summarization," in *2021 44th Int. ACM SIGIR Con. Res. and Dev. in Info. Retri.*, Canada, Jul. 11–15, 2021, pp. 2146–2150. doi: 10.1145/3404835.3463037.
[10] F. Ebrahimi and A. Mahmoud, "Unsupervised summarization of privacy concerns in mobile application reviews," in *2022 37th IEEE/ACM Int. Conf. Auto. Soft. Eng.*, Rochester, MI, USA, Oct. 10, 2022, pp. 1–12. doi: 10.1145/3551349.3561155.
[11] R. Sultana and S. Sarker, "App review mining and summarization," *Int. J. Comp. App.*, vol. 975, pp. 8887, Apr. 2018. doi: 10.3390/electronics12051258.
[12] H. Malik, E. M. Shakshuki, and W. S. Yoo, "Comparing mobile apps by identifying 'Hot' features," *Future Gen. Comp. Syst.*, vol. 107, pp. 659–669, 2020. doi: 10.1016/j.future.2018.02.008.
[13] M. Assi, S. Hassan, Y. Tian, and Y. Zou, "FeatCompare: Feature comparison for competing mobile apps leveraging user reviews," *Emp. Soft. Eng.*, vol. 26, pp. 94, Sep. 2021. doi: 10.1007/s10664-021-09988-y.
[14] Y. Li, B. Jia, Y. Guo, and X. Chen, "Mining user reviews for mobile app comparisons," in *Proc. ACM on Interact. Mob. Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–15, Sep. 2017. doi: 10.1145/3130935.
[15] T. P. Liang, X. Li, C. T. Yang, and M. Wang, "What in consumer reviews affects the sales of mobile apps: A multifacet sentiment analysis approach," *Int. J. Electron. Commer.*, vol. 20, no. 2, pp. 236–260, Dec. 2015. doi: 10.1080/10864415.2016.1087823.
[16] R. A. Masrury, A. Alamsyah, and F. Fannisa, "Analyzing tourism mobile applications perceived quality using sentiment analysis and topic modeling," Present at the 2019 7th ICoICT, Kuala Lumpur, Malaysia, Jul. 24–26, 2019, pp. 1–6. doi: 10.1109/ICoICT.2019.8835255.
[17] M. E. Permana, H. Ramadhan, I. Budi, A. B. Santoso, and P. K. Putra, "Sentiment analysis and topic detection of mobile banking application review," in *2020 Fifth ICIC*, Nov. 3–4, 2020, pp. 1–6. doi: 10.1109/ICIC50835.2020.9288616.
[18] M. R. Abdul Rahim, S. Abdul-Rahman, and Y. Mahmud, "Customers' opinions on mobile telecommunication services in Malaysia using sentiment analysis," *Int. J. Adv. Comp. Sci. Appl.*, 2021. doi: 10.14569/issn.2156-5570.
[19] M. J. Kim, "Analyzing the trend of wearable keywords using text-mining methodology," *J. Dig. Conv.*, pp. 181–190, Sep. 2021. doi: 10.14400/JDC.2020.18.9.181.
[20] J. Zhou and M. Zhou, "Sentiment analysis of elderly wearable device users based on text mining," in *2021 AHFE Virt. Conf. Usability and User Exp.*, USA, Springer International Publishing, Jul. 25–29, 2021, pp. 360–365. doi: 10.1007/978-3-030-80091-8_42.
[21] S. Asthana, A. Megahed, and R. Strong, "A recommendation system for proactive health monitoring using IoT and wearable technologies," in *2017 IEEE Int. Conf*, Hawaii, USA, Jun. 25, 2017, pp. 14–21. doi: 10.1109/AIMS.2017.11.
[22] E. Vermetten *et al.*, "Using VR-based interventions, wearable technology, and text mining to improve military and Veteran mental health," *J. Milit. Vet. Fam. Health*, pp. 26–35, Mar. 2020. doi: 10.3138/jmvfh.2019-0033.
[23] G. A. Vermetten, M. W. Jaspers, M. P. Schijven, and L. W. Dusseljee-Peute, "Mobile health for older adult patients: Using an aging barriers framework to classify usability problems," *Int. J. Med. Inform.*, pp. 68–77, Jan. 15, 2019. doi: 10.1016/j.ijmedinf.2019.01.006.
[24] B. A. Kumar and M. S. Goundar, "Usability heuristics for mobile learning applications," *Educ. Inform. Tech.*, pp. 1819–1833, Jan. 17, 2019. doi: 10.1016/j.ijmedinf.2019.01.006.
[25] J. Nielsen, "Usability engineering, Morgan Kaufmann, 1994.
[26] I. Delikostidis, T. Fechner, H. Fritze, A. M. AbdelMouty, and C. Kray, "Evaluating mobile applications in virtual environments: A survey," *Int. J. Mob. Hum. Comput.*, pp. 1–19, Oct. 2013. doi: 10.4018/ijmhci.2013100101.

[27] K. Moumane, A. Idri, and A. Abran, "Usability evaluation of mobile applications using ISO, 9241 and ISO 25062 standards," *SpringerPlus*, vol. 29, pp. 1–15, Apr. 2016. doi: 10.1186/s40064-016-2171-z.
[28] H. K. Flora, X. Wang, and S. V. Chande, "An investigation on the characteristics of mobile applications: A survey study"," *Int. J. Modern Educ. Comput. Sci.*, vol. 6, no. 11, pp. 21–27, Nov. 2014. doi: 10.5815/ijitcs.2014.11.03.
[29] A. Méndez, L. C. U. Quesada, and C. M. Jenkins, "Automated testing of mobile applications: A systematic map and review," in *10th Int. Conf. Web Info. Syst. Tech.*, Amman, Jordan, Jul. 14–15, 2021.
[30] R. Harrison, D. Flood, and D. Duce, "Usability of mobile applications: Literature review and rationale for a new usability model," *J. Interact Sci.*, vol. 1, pp. 1–16, May 2013. doi: 10.1186/2194-0827-1-1.
[31] A. D. Sorbo *et al.*, "What would users change in my app? Summarizing app reviews for recommending software changes," in *2016 24th ACM SIGSOFT Int. Symp. Foundations Soft. Eng.*, Seattle, WA, USA, Nov. 13–18, 2016, pp. 499–510. doi: 10.1145/2950290.2950299.
[32] N. Jha and A. Mahmoud, "Using frame semantics for classifying and summarizing application store reviews," *Emp. Soft. Eng.*, vol. 23, pp. 3734–3767, Mar. 2018. doi: 10.1007/s10664-018-9605-x.
[33] M. Lu and P. Liang, "Automatic classification of non-functional requirements from augmented app user reviews," in *21st Int. Conf. Eval. and Asses. Soft. Eng*, Jun. 15, 2017, pp. 344–353. doi: 10.1145/3084226.3084241.
[34] S. Hassan, C. Tantithamthavorn, C. P. Bezemer, and A. E. Hassan, "Studying the dialogue between users and developers of free apps in the google play store," *Emp. Soft. Eng.*, vol. 23, pp. 1275–1312, Sep. 2018. doi: 10.1007/s10664-017-9538-9.
[35] E. Noei, F. Zhang, and Y. Zou, "Too many user-reviews! What should app developers look at first?" *IEEE Trans. Soft. Eng.*, vol. 47, no. 2, pp. 367–378, Jan. 2019. doi: 10.1109/TSE.2019.2893171.
[36] A. K. Alssayh, "Mobile Hajj guide for Malaysian pilgrims," Ph.D. dissertation, Universiti Utara Malaysia, Malaysia, 2009.
[37] M. M. Anad, "A Mobile Application T0 Guide Hajj Pilgrims," Masters thesis, Universiti Utara Malaysia, Malaysia, 2009.
[38] A. A. Owaidah, "Hajj crowd management via a mobile augmented reality application: A case of The Hajj event, Saudi Arabia," Ph.D. dissertation, Univ. of Glasgow, UK, 2015.
[39] M. Sakina, "The Four Basic Pillars of Hajj's Mobile System, " Masters thesis, Universiti Utara Malaysia, Malaysia, 2011.
[40] T. Mantoro, M. Akhtaruzzaman, M. Mahmud, and M. A. Ayu, " Design and development of an interactive monitoring system for Pilgrims in congregation of Hajj ritual," *J. Conv. Inf. Technol.*, vol. 10, no. 1, pp. 28–57, Jan. 2015.
[41] M. A. Almasry, *The Use of Augmented Reality as Assistive Services for Pilgrims*. KSA, 2011.
[42] A. S. A. Al-Aidaroos, A. N. Zulkifli, and R. C. Mat, "Development of mobile dua and zikr for Hajj (MDZ4H)," *TELKOMNIKA Indonesian J. Elec. Eng.*, vol. 11, no. 5, pp. 2723–2730, May 2013. doi: 10.11591/telkomnika.v11i5.2509.
[43] A. Ahmad, M. Abdur Rahman, I. Afyouni, F. Rehman, B. Sadiq and M. R. Wahiddin, "Towards a mobile and context-aware framework from crowdsourced data," in *Proc. 5th ICT4M*, Kuching, Malaysia, Nov. 17–18, 2014, pp. 1–6. doi: 10.1109/ICT4M.2014.7020672.
[44] A. M. Zeki, H. Alsafi, R. M. Nassr, and T. Mantoro, "A mobile dictionary for pilgrims," in *2012 Int. Conf. Info. Tech. E-Serv.*, Sousse, Tunisia, 2012, pp. 1–5. doi: 10.1109/ICITeS.2012.6216620.
[45] A. Amro and Q. Nijem, *"Pilgrims" Hajj "Tracking System (e-Mutawwif)"*. Contemp. Eng. Sci. Med., KSA, pp. 437–446, 2012.
[46] F. Hamhoum and C. Kray, "Supporting pilgrims in navigating densely crowded religious sites," *Pers. and Ubiq. Comput.*, vol. 16, pp. 1013–1023, Oct. 2011. doi: 10.1007/s00779-011-0461-6.
[47] H. H. Mohamed, "M-Umrah: An Android-based application to help pilgrims in performing Umrah," in *2013 Int. Con. on Adv. Comp. Sci. App. Tech.*, Dec. 23–24, 2013, pp. 385–389. doi: 10.1109/ACSAT.2013.82.
[48] L. Zhou, Z. Zhang, L. Zhao, and P. Yang, "Attention-based BiLSTM models for personality recognition from user-generated content," *Info. Sci.*, vol. 596, pp. 460–471, Jun. 2022. doi: 10.1016/j.ins.2022.03.038.

[49] Z. Zhang, J. Guo, H. Zhang, L. Zhou, and M. Wang, "Product selection based on sentiment analysis of online reviews: An intuitionistic fuzzy TODIM method," *Compl. Intell. Syst.*, vol. 8, pp. 3349–3362, Feb. 2022. doi: 10.1007/s40747-022-00678-w.

[50] L. Zhou, L. Tang, and Z. Zhang, "Extracting and ranking product features in consumer reviews based on evidence theory," *J. Amb. Intell. Hum. Comput.*, vol. 14, pp. 9973–9983, Jan. 2023. doi: 10.1007/s12652-021-03664-1.

[51] M. Mohandes, "Pilgrim tracking and identification using the mobile phone," in *2011 IEEE 15th ISCE*, Jun. 14–17, 2011, pp. 196–199. doi: 10.1109/ISCE.2011.5973812.

[52] M. A. Abdelazeez and A. Shaout, "Pilgrim communication using mobile phones," *J. Image Graph.*, vol. 4, no. 1, 2016. doi: 10.18178/joig.4.1.59-62.

[53] A. Shaout and S. Khan, "ALHAJJ â€HAJJ APP FOR IOS," *IIUM Eng. J.*, vol. 17, no. 1, pp. 1–27, 2016. doi: 10.31436/iiumej.v17i1.528.

[54] A. S. Alqahtani *et al.*, "Pilot use of a novel smartphone application to track traveller health behaviour and collect infectious disease data during a mass gathering: Hajj pilgrimage 2014," *J. Epidemiol. Glob. Health*, vol. 6, no. 3, pp. 147–155, 2016. doi: 10.1016/j.jegh.2015.07.005.

[55] E. A. Khan and M. K. Y. Shambour, "An analytical study of mobile applications for Hajj and Umrah services," *Appl. Comput. Inform.*, vol. 14, no. 1, pp. 37–47, 2018. doi: 10.1016/j.aci.2017.05.004.

[56] S. Nusinovici *et al.*, "Logistic regression was as good as machine learning for predicting major chronic diseases," *J. Clin. Epidemiol.*, vol. 122, pp. 56–69, 2020. doi: 10.1016/j.jclinepi.2020.03.002.

[57] X. Zou, Y. Hu, Z. Tian, and K. Shen, "Logistic regression model optimization and case analysis," Present at the 2019 IEEE 7th ICCSNT, 2019, pp. 135–139. doi: 10.1109/ICCSNT47585.2019.8962457.

[58] G. Battineni, N. Chintalapudi, and F. Amenta, "Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM)," *Inform. Med. Unlocked*, vol. 16, pp. 100200, 2019. doi: 10.1016/j.imu.2019.100200.

[59] S. Ghosh, A. Dasgupta, and A. Swetapadma, "A study on support vector machine based linear and non-linear pattern classification," in *2019 Int. Conf. Inf. Soc. Sci.*, Manila, Philippines, IEEE, Jul. 12–14, 2019, pp. 24–28. doi: 10.1109/ISS1.2019.8908018.

[60] M. M. Nishat, F. Faisal, T. Hasan, M. F. B. Karim, Z. Islam and M. R. K. Shagor, "An investigative approach to employ support vector classifier as a potential detector of brain cancer from MRI dataset," in *2021 ICECIT*, Hangzhou, China, Mar. 5–7, 2021, pp. 1–4. doi: 10.1109/ISS1.2019.8908018.

[61] M. J. Hasan, S. Mahbub, M. S. Alom, and M. A. Nasim, "Rice disease identification and classification by integrating support vector machine with deep convolutional neural network," in *2019 1st ICASERT*, Dhaka, Bangladesh, May 3–5, 2019, pp. 1–6. doi: 10.1109/ICASERT.2019.8934568.

[62] H. Jelodar, Y. Wang, R. Orji, and S. Huang, "Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 10, pp. 2733–2742, 2020. doi: 10.1109/JBHI.2020.3001216.

[63] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning," *IEEE Access*, vol. 8, pp. 23522–23530, 2020. doi: 10.1109/ACCESS.2020.2969854.

[64] L. Zhou and X. Bian, "Improved text sentiment classification method based on BiGRU-attention," *Int. J. Phys. Conf. Ser.*, vol. 1345, no. 3, pp. 32097, 2019. doi: 10.1088/1742-6596/1345/3/032097.

[65] J. Salminen, M. Hopf, S. A. Chowdhury, S. G. Jung, H. Almerekhi and B. J. Jansen, "Developing an online hate classifier for multiple social media platforms," *Hum. Centr. Comp. Inf. Sci.*, vol. 10, pp. 1–34, 2020. doi: 10.1186/s13673-019-0205-6.

[66] M. A. Al-Ajlan and M. Ykhlef, "Deep learning algorithm for cyberbullying detection," *Int. J. Adv. Comp. Sci. Appl.*, vol. 9, no. 9, 2018. doi: 10.14569/issn.2156-5570.

[67] J. Kocoń, A. Figas, M. Gruza, D. Puchalska, T. Kajdanowicz and P. Kazienko, "Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach," *Inf. Proc. Manage.*, vol. 58, no. 5, pp. 102643, 2021. doi: 10.1016/j.ipm.2021.102643.

[68] M. U. Salur and I. Aydin, "A novel hybrid deep learning model for sentiment classification," *IEEE Access*, vol. 8, pp. 58080–58093, 2020. doi: 10.1109/ACCESS.2020.2982538.

[69] T. Zheng, F. Wu, R. Law, Q. Qiu, and R. Wu, "Identifying unreliable online hospitality reviews with biased user-given ratings: A deep learning forecasting approach," *Int. J. Hosp. Manage.*, vol. 92, pp. 102658, 2021. doi: 10.1016/j.ijhm.2020.102658.